Predicting Longevity Using Urban Greenspace

Problem Statement

Access to urban greenspace is becoming increasingly important as cities continue to grow, and studies have linked access to greenspace (e.g. parks, gardens) to mental well-being (Thompson et al. 2016). Over the past few years, London's local government has been making a push for "greening London"; the London Plan 2021 outlines an integrated environmental strategy for increasing urban greenspace. With the current population of London at around 9,648,000, and a trend of ~1-1.5% population increase per year over the past few years (United Nations World Population Prospects), the importance of urban design has become an important consideration for the city's urban planning commissions. The focus of this project was to develop a model to predict which factors, including access to urban greenspace, are most important for predicting longevity within the London Wards.

Data Wrangling

I used data from the London Ward Well-Being Scores dataset, collected by the Greater London Authority over the period of 2009-2013. This dataset includes information on life-expectancy, as well as 12 different well-being indicators such as access to greenspace and public transport, childhood obesity rates, crime rates, and unemployment rates. I merged a second well-being dataset from the Greater London Authority that uses a different metric (total area) for urban greenspace so I could assess whether different greenspace metrics affect the model.

I used fuzzy string matching to merge the two datasets and checked for missing values; the cleaned merged dataset contained 3,275 rows and 16 columns.

Exploratory Data Analysis and Preprocessing

I used panda profiling for a first examination of the data, as well as heatmaps and pairplots (seaborn) to visualize relationships between variables. The variables that have the strongest correlation with life expectancy are unemployment rate (-0.63), number of dependent children in out-of-work households (-0.69), school absences (-0.55), incapacity benefit (-0.57), and GCSE points (+0.55).

These correlations all make sense intuitively; the variables with negative correlation are all representative of hardship. Specifically, the "dependent children" metric represents the % of children living in out-of-work households, and "incapacity benefits" represents the disability claimant rate for that ward. GCSE_points, the positive correlate, is an indication of dependent children doing well in school.

I used ColumnTransformer to create a pipeline that performed one-hot-encoding on categorical variables and standard scaler on numerical variables. Since I was interested in the effect of greenspace on well-being (life expectancy), I created a binned ordinal metric for greenspace that designated it as low, medium, or high based on the total area of greenspace in each region (Borough) of London.

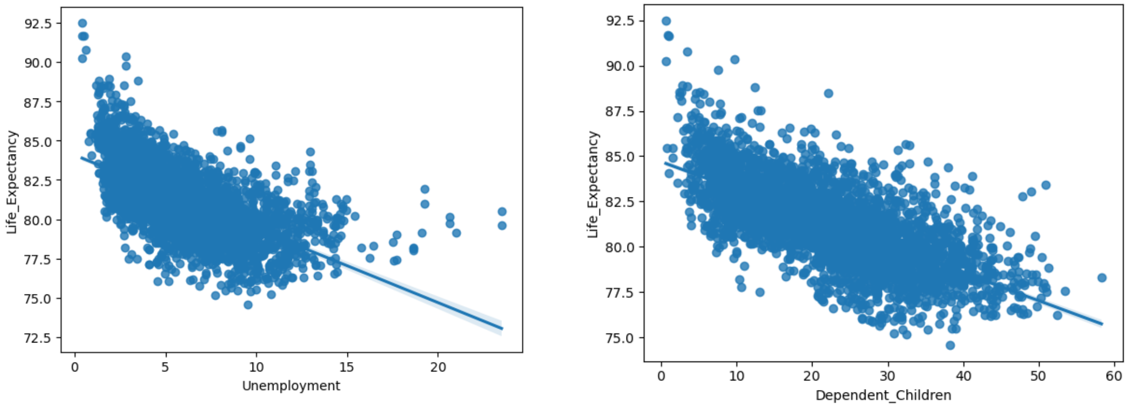

Figure 1: Unemployment rate and the number of dependent children living in out-of-work households versus life expectancy.

The figure above shows the negative correlation between life expectancy and two features of interest, namely unemployment rate and the number of dependent children living in out-of-work households.

Model Selection

This is a regression problem in supervised machine learning. I tested the following four regression models:

-- Linear Regression
-- Random Forest
-- Gradient Boosting Regressor
-- Support Vector Regressor (SVR)

I used r-squared and MAE to determine that the gradient boosting model performed the best, and then used GridSearchCV to tune hyperparameters using k-fold cross-validation. The gradient boosting model was successful at predicting longevity within the London Wards to within ~1 year, with fairly high predictive capacity (R-squared = 0.770).

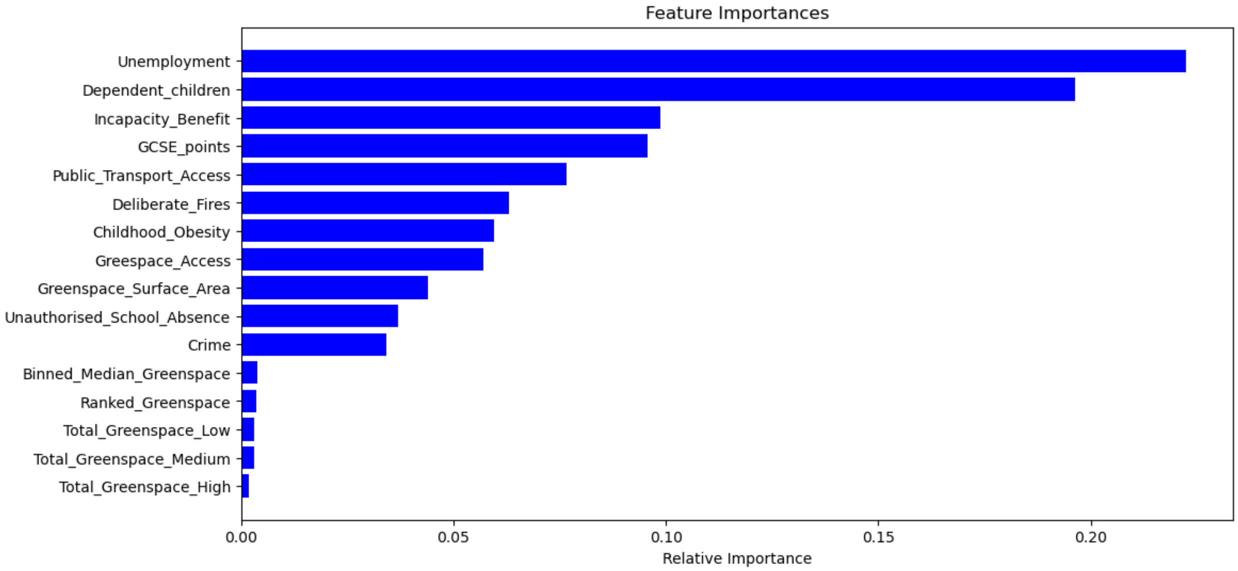

Figure 2: Feature importance for the scaled features in the gradient boosting model.

As can be seen in the figure above, unemployment and the number of dependent of dependent children living in out-of-work households were by far the strongest predictors of longevity. However, the model also shows that there are many factors that play a role in predicting well-being (measured as longevity), including both metrics of urban greenspace.

Key Findings:

1. The most important predictor of longevity is unemployment rate.

Unemployment rate indicates the percentage of working-age residents claiming unemployment benefits in that ward. This, unsurprisingly, suggests that there is a high degree of stress related to being unemployed that negatively impacts longevity. Investigating and addressing underlying causes of unemployment is necessary for supporting the well-being of London residents.

2. Unauthorized school absences, the number of dependent children living in an out-of-work household, and incapacity benefit were also important in predicting longevity.

These metrics are all representative of hardship; specifically, the "dependent children" metric represents the % of children living in out-of-work households, and "incapacity benefits" represents the disability claimant rate for that ward.

3. Access to public greenspace areas does provide some positive benefit.

While access to public greenspace was not a strong predictor of longevity, it does show some importance in the model. Looking at the original data, we can see a positive correlation between greenspace access and longevity. This suggests a positive benefit for people living in areas with more access to parks or other green areas; this may be due to the impact on lifestyle that this access affords, or possibly is related to the wealth of that area.

Future Research:

The final model did well at predicting longevity for London residents, with 11 features showing the strongest predictive importance. However, this was a small study on a specific city, and it is not clear how well this model would generalize to other cities. One interesting question to address would be whether this model works well for similar sized cities, and then whether it could be generalized to much smaller cities. Ideally, I would be able to train the model on cities of varying size, and include population size or density as a parameter.

Additionally, this model made me think about other ways to characterize and analyze the importance of urban greenspace to well-being. While access to urban greenspace and the total amount of greenspace in a city are important, they were far from the most important feature in the model. Measuring “well-being” is no easy task; here, I used longevity as a measure of well-being, but other metrics could be explored as well (e.g., rates of depression or other health indicators).